# 可驱动的泛化人头神经辐射场

王月[1]

[1]（中国科学技术大学大数据学院, 合肥 230026）

**摘 要：**近年来，随着计算机视觉领域的快速发展，数字人的概念引起社会各界的广泛关注，高保真的人体、人头和人手的建模都得到了深入的研究。本文关注头部建模，基于神经辐射场提出一种可泛化的人头模型，结合人脸识别网络和人脸三维形变模型，将头部模型参数化，因此可以直接控制生成图像的身份和表情语义属性，并且支持自由编辑图像的渲染姿态。为了提高神经辐射场的渲染速度，我们将传统的体渲染改为体渲染结合二维神经渲染的方式，在保留渲染图像质量的同时在 Tesla V100 GPU 上达到 15 帧/秒的渲染速度。通过采集大量的头部 RGB 图像数据参与训练，模型可以生成高保真的渲染图像，并且在测试集上也有逼真的拟合结果，可以泛化到未曾参与训练的新的身份和表情语义。得益于神经辐射场对三维几何场景的隐式表示能力，模型的渲染结果具有多视角一致性，在新视角合成、表情迁移、驱动等方面有多种用途。

**关键词：**神经辐射场；人脸参数化模型；泛化；驱动；语义解耦
文献标志码：A　　中图分类号:TP ***

# Drivable generalized head neural radiance field

WANG Yue[1]

[1] (School of Data Science, University of Science and Technology of China, Hefei 230026, China)

**Abstract:** In recent years, with the rapid development of computer vision, the concept of digital human has attracted wide attention from all walks of life, and the modeling of high-fidelity human body, head and hand has been deeply studied. This paper focus on head modeling and propose a generalized head model based on neural radiance field, which is parameterized by face recognition network and 3D face morphable model, therefore, it can directly control the semantic attributes such as identity and expression of the generated image, and support freely modifying the rendering pose of the image. In order to improve the rendering speed of neural radiance field, this paper use the combination of the volume rendering and two-dimensional neural rendering to replace the traditional pure volume rendering, which can speed up the rendering process while preserving image quality. The head model can render images with the speed of 15 frames per second on the Tesla V100 GPU. By collecting a large amount of head RGB images data to participate in training stage, the model can generate high-fidelity rendering images, and also have realistic fitting results on the test set, it can be generalized to new identity and expression that have not been trained. Thanks to the ability of implicit representation of 3D geometric scene by neural radiance field, the rendering results of the model has multi-view consistency, and has many uses such as novel view synthesis, expression transfer, driving and so on.

**Key words:** neural radiance field (NeRF); face parametric model; generalization; driving; semantic disentanglement

三维人脸/头部表示是近年来计算机视觉和计算机图形学领域的热点问题之一，在增强现实（Augmented Reality, AR）/虚拟现实（Virtual Reality, VR）、数字游戏、电影制作等方面均有广泛应用。如何高度保真地重建出视频或图像中的人头模型是一个非常具有挑战性的研究课题。

基于人头模型可以嵌入到低维空间的假设，参数化语义人头模型如 blendshape，在很长时间内被众多学者加以研究和改进。Blendshape 人头模型是一个以不同面部表情的线性/双线性组合的头部模型，具有语义参数化性质，用户可以通过组合系数决定面部表情。同时，blendshape 构建了一个合理的形状空间，帮助用户实现自定义的个性化人脸编辑。在此基础上，具有泛化性的语义人头模型如 FaceWarehouse[1]旨在用不同的表情对不同的对象进行建模，但忽略了可能存在的几何和纹理细节。为了构建表达能力更强的 blendshape 模型，传统的基于网格的方法通常采用基于多线性张量的三维形变模型（3D Morphable Model, 3DMM）[2]，但这种建模方式通常会忽略牙齿、头发等非人脸部分，此外，由于网格渲染的分辨率限制，模型很难表达出皱纹等高频的细节信息，且网格渲染并不可微，要使用 RGB 图像监督训练必须采用近似可微渲染技术来缓解不可微问题。

近年来，随着深度学习的飞速发展，二维生成对抗网络（generative adversarial networks, GANs）[3,4]能够在不需要三维模型的情况下直接渲染高质量的人脸图像，给人头图像生成质量带来了巨大的提升。但 GAN 不涉及三维几何模型，缺乏语义信息，导致表情属性无法被轻易控制。因此一些工作[5,6]进一步考虑加入解耦约束实现用户自定义的人脸图像渲染。然而，由于没有显示的三维模型约束，这些生成方法在不同视角下呈现的渲染结果往往不具有多视角一致性。

2020 年，Mildenhall 等人[7]提出用神经辐射场（Neural Radiance Fields, NeRF）表示三维场景，这种表示方法可以合成逼真的渲染图像，因此迅速成为图像生成领域引人注目的研究方法，一大批基于神经辐射场的人脸/人头图像生成工作应运而生。一些学者也考虑将 GAN 和 NeRF 结合在一起[8–13]以生成高保真的人脸图像，但这种生成模型仍然与基础的 GAN 模型有同样的缺点，将身份、表情和外观全部耦合在一起，无法进行语义编辑。HeadNeRF[14]提出利用 3DMM 解耦不同的语义属性建立头部参数化模型，通过海量的高清单人图像训练模型的泛化能力，生成了可语义编辑的人头神经辐射场。

NeRF 实际上是对三维几何场景的一种隐式编码，可以看成是一个带有纹理的网格，具有良好的多视角一致性，可以自然的进行可微渲染，并且除了二维图像外不需要任何额外的三维真实数据作为标签，通过渲染结果与输入图像之间的误差约束，就可以实现端到端的自监督训练过程。但是 NeRF 的场景优化建立在密集多视角图像的基础上，不同场景之间无法共享相同的先验知识，因此每个场景或对象必须每次单独优化，不具有良好的泛化性能。为了解决这一问题，一些学者在 NeRF 的训练中引入局部特征，通过二维卷积神经网络[15,16]或者 Transformer 网络[17]提取输入图像中的特征信息，作为额外输入加入到 NeRF 的训练中，生成了具有泛化性的模型，提高了 NeRF 的模型表达能力。

基于上述观察，为了重建一个具有语义编辑功能的可泛化人头模型，本文将 NeRF 应用到人体头部的表示上，提出可驱动的泛化人头神经辐射场模型。我们的模型继承了 NeRF 的优点，不仅可以生成高保真的人头图像，在多视角一致性方面也有显著表现。由于 NeRF 本身支持自由变换用于渲染的相机视角，因此我们的模型也自然地支持人头图像的姿态编辑。此外，模型只需要二维图像作为输入，因为渲染阶段可微，所以仍然是自监督训练，不需要任何额外的三维几何监督信号。通过精心设计的网络结构以及损失函数，结合身份和表情特征，在大量训练数据的支撑下，模型的泛化性能也得到保证。具体来说，我们收集并处理了一个包含多人的单目动态视频数据集，通过拟合 3DMM 模型[18]得到表情系数作为特征之一，利用人脸识别网络提取身份特征作为特征之二，将这两个特征作为 NeRF 的额外输入优化人头模型的表示，通过多身份多表情数据的多轮训练后，模型成功的解耦了身份和表情，实现了可驱动功能。

进一步地，我们将 NeRF 的体渲染与二维神经渲染相结合，在模型推理阶段将渲染速度提高了数倍，在 Tesla V100 GPU 上达到 15 帧/秒。与 GIRAFFE[11]和 StyleNeRF[10]类似，这种从粗到细（coarse-to-fine）的策略在不牺牲渲染质量的前提下显著加快了渲染速度。得益于良好的解耦表示、快速的推理渲染和高保真的生成结果，我们的模型可以有各种应用，如单张人脸图像的新视角合成、表情迁移、姿态驱动等。总的来说，我们建立的是一个基于神经辐射场的可驱动的具有泛化性的人头部模型。

## 1 模型的建立与表示

本文旨在建立一个可驱动的泛化人头模型，该模型不仅可以通过语义编辑实现头部驱动，而且在新的数据集上也具有良好的拟合效果。为实现这一目的，本文将神经辐射场作为新的三维代理，代替传统的人脸参数化模型，并结合人脸识别网络，提出了一个新的可控制身份、表情、渲染视角的泛化模型。与之前

基于三维网格的生成方法不同，采用 NeRF 的加速变体作为统一的三维代理，可以直接控制渲染结果的相机视角，并在 GPU 上实现高保真可驱动的人头图像。为了训练模型，我们收集并处理了一个单目动态视频数据集，包含多身份和多表情，为训练提供了大量数据，得益于此，该模型具有一定的泛化能力。同时，本文设计了合适的网络结构和损失函数，使训练后的模型能够控制身份和表情属性，从而实现逼真的驱动效果。

## 1.1 神经辐射场与人脸参数化模型的回顾

### 1.1.1 神经辐射场

这部分将简单回顾 NeRF 表示[7]。NeRF 将场景编码成一个与颜色和密度相关的连续体素辐射场 f，具体来说，对于一个三维空间点 $\boldsymbol{x}\in\mathbb{R}^3$ 和一个视角方向 $\boldsymbol{d}\in\mathbb{R}^3$，在经过位置编码 $\gamma(*)$ 作用后通过函数 f 映射到一个可微的体密度 $\sigma$ 和一个 RGB 颜色 $\boldsymbol{c}$。

$$f_\theta:(\gamma(\boldsymbol{x}),\gamma(\boldsymbol{d}))\mapsto(\sigma,\boldsymbol{c}) \quad (1.1)$$

然后，这个体素辐射场可以利用下面的公式通过可微渲染生成二维图像：

$$\hat{C}(\boldsymbol{r})=\int_{t_n}^{t_f}T(t)\sigma(t)\boldsymbol{c}(t)dt \quad (1.2)$$

其中，$T(t)=\exp\left(-\int_{t_n}^{t}\sigma(s)ds\right)$ 表示射线从 $t_n$ 到 $t$ 的累计透明度，即射线从 $t_n$ 到 $t$ 不撞击任何其他粒子的概率。假设目标视角的相机参数为 $\boldsymbol{P}$，那么一条从相机中心射出的光线可以将其表示为 $\boldsymbol{r}(t)=\boldsymbol{o}+t\boldsymbol{d}$，其中射线原点 $\boldsymbol{o}\in\mathbb{R}^3$ 为相机中心，射线的单位方向向量为 $\boldsymbol{d}\in\mathbb{R}^3$。(1.2)式中的积分沿着射线 $\boldsymbol{r}$ 在预先设定的深度边界 $\left[t_n,\ \ t_f\right]$ 内计算，在实际实现时，射线取相机中心到图像上每个像素的连线，该积分则被近似为射线上每个采样点的数值积分。

对于相机参数为 $\boldsymbol{P}$ 的目标视图，由相机中心发出的一条射线记为 $\boldsymbol{r}$，在这条射线上利用(1.2)式渲染得到的像素值 $\hat{\boldsymbol{C}}(\boldsymbol{r})$ 可以与真实图像上相应的像素值 $\boldsymbol{C}(\boldsymbol{r})$ 进行比较，由此可以写出 NeRF 的渲染误差如下式所示：

$$\mathcal{L}=\sum_{\boldsymbol{r}\in\mathcal{R}(\boldsymbol{P})}\left\|\hat{C}(\boldsymbol{r})-\boldsymbol{C}(\boldsymbol{r})\right\|_2^2 \quad (1.3)$$

其中 $\mathcal{R}(\boldsymbol{P})$ 为相机参数为 $\boldsymbol{P}$ 时所有由相机中心发出的射线形成的集合。

NeRF 表示在新视角合成方面的工作中取得了非常理想的效果，与经典多视角立体匹配方法一样[19,20]，它是一种基于优化的方法，唯一的优化信息来源于几何一致性。而不同场景的几何信息无法共享[21]，因此它必须在每个场景下单独优化，不具有良好的泛化性能。当场景不同时，训练模型将需要耗费很多时间。并且一旦视角稀少，无法利用现实世界中任何先验知识重建出物体的完整形状[22–24]。要想在有限视角下重建出具有泛化性能的 NeRF 模型，可以考虑加入局部特征[15–17]增强模型的泛化能力。

### 1.1.2 人脸参数化模型

3D 形变模型[18](3D Morphable model, 3DMM)是使用最为广泛的一类三维人脸参数化模型，它将空间中三维人脸的几何和反照率编码到低维子空间中。具体来说，3DMM 模型用主成分分析法(Principle Component Analysis, PCA)描述了三维人脸几何形状 $\boldsymbol{S}$ 和反照率 $\boldsymbol{b}$：

$$\boldsymbol{S}=\bar{\boldsymbol{S}}+\boldsymbol{\alpha}\boldsymbol{A}_{id}+\boldsymbol{\beta}\boldsymbol{A}_{exp} \quad (1.4)$$

$$\boldsymbol{b}=\bar{\boldsymbol{b}}+\boldsymbol{\delta}\boldsymbol{A}_{alb} \quad (1.5)$$

其中 $\bar{\boldsymbol{S}}$ 和 $\bar{\boldsymbol{b}}$ 分别表示平均人脸的形状和反照率，$\boldsymbol{A}_{id},\boldsymbol{A}_{alb}$ 分别表示从一组具有中性表情的带纹理的三维网格中提取的主轴，$\boldsymbol{A}_{exp}$ 表示在每个个体带表情的网格和中性表情网格之间的偏移量上训练的主轴，$\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\delta}$ 则是表征特定三维人脸模型的相应系数向量。为了多样性和互补性，本文使用 Basel Face Model (BFM)[25] 生成三维人脸的形状和反照率，用 FaceWarehouse[1]生成表情基，特别地，本文使用的所有表情系数的维度为 46。

## 1.2 模型表示

我们认为头部图像的几何形状主要由身份和表情相关的隐编码控制，这与 3DMM 的底层逻辑是一致的。具体地说，我们将身份和表情视为每个对象的特征信息，并以此作为 NeRF 的额外输入，以保证模型在结构上具有泛化性。为了表征表情属性，将拟合 3DMM 模型得到的表情系数作为表情隐编码，记为 $\boldsymbol{\beta}$。考虑到身份信息是每个人独一无二的，不会随着表情或光照等其他情况的改变而变化，这与人脸识别的目的不谋而合，因此本文的身份隐编码考虑使用人脸识别网络提取的相关特征信息，为此我们找到了目前开源的准确度最高的人脸识别网络 AdaFace[26]，用其提供的预训练模型从人头图像中提取人脸特征作为身份隐编码，记为 $\boldsymbol{z}_{id}$。通过加入表情编码和身份编码作为条件输入，可将(1.1)式中基于 MLP 的隐函数 $f_\theta$ 改写成

下式，以建立本文提出的模型：

$$f_\theta:(\gamma(\boldsymbol{x}),\gamma(\boldsymbol{d}),\boldsymbol{\beta},\boldsymbol{z_{id}})\mapsto(\sigma,\boldsymbol{F}) \tag{1.6}$$

其中 $\theta$ 表示网络可优化的参数，$f_\theta$ 表示的网络结构以及完整模型的整体框架如图 1 所示。这里 $\gamma(*)$ 为预先定义的位置编码函数，与 NeRF 原文[7]所使用的位置编码设置相同。

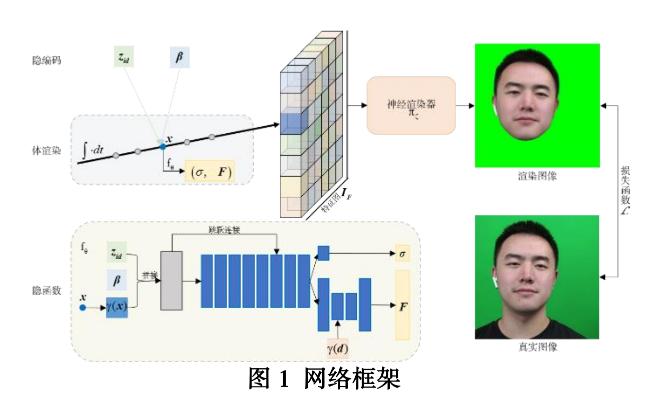


**图 1 网络框架**

Fig.1 Network Framework

在图 1 中，$\boldsymbol{x}\in\mathbb{R}^3$ 是射线上的一个三维采样点。与之前的工作[10-12]类似，我们并不直接预测采样点 $\boldsymbol{x}$ 的 RGB，而是预测一个高维的特征向量 $\boldsymbol{F}(\boldsymbol{x})\in\mathbb{R}^{512}$，用体渲染结合神经渲染的方式代替单一的体渲染模块。在这里之所以不采用传统 NeRF 的方式预测三通道颜色值，是因为这种方式在后续体渲染阶段需要对每个采样点的颜色值计算数值积分，而为了能够渲染出质量较高的图像则需要在射线上采样大量的三维空间点，这将在体渲染阶段消耗大量的计算资源和时间。如果预测特征向量的话，就可以在体渲染阶段首先将场景渲染到一个较低分辨率的特征图，再经过神经渲染器处理特征图并输出最终的 RGB 图像，这种方式可以对算法进行加速并且节省计算资源。具体来说，我们将位置编码函数作用在采样点 $\boldsymbol{x}$ 上得到 $\gamma(\boldsymbol{x})$ 后，与身份隐编码 $\boldsymbol{z_{id}}$ 以及表情隐编码 $\boldsymbol{\beta}$ 拼接在一起后作为整个网络的输入，通过若干层 MLP 输出体密度 $\sigma$ 和中间特征，然后再将中间特征和经过位置编码函数作用后的视角方向 $\gamma(\boldsymbol{d})$ 拼接后再次通过 MLP，进一步预测特征向量 $\boldsymbol{F}(\boldsymbol{x})$。这样的网络结构使得密度场的预测只与身份和表情隐编码相关，而不受视角方向的影响，视角方向的变化只会影响特征预测的结果，进而影响渲染图像的像素 RGB 数值。这与现实世界所呈现的物理原理是一致的，密度场表示的是物体与场景的几何信息，并不会随着观察方向的改变发生变化，而在不同视角下看到的彩色成像结果应该因为光线等因素略有不同。

根据以上描述，本文所建立模型的体渲染阶段会得到一个低分辨率的特征图 $\boldsymbol{I_F}\in\mathbb{R}^{512\times32\times32}$，参考 NeRF 中体渲染公式，可以写出本模型中体渲染阶段的公式：

$$\boldsymbol{I_F}(\boldsymbol{r})=\int_0^\infty w(t)\cdot\boldsymbol{F}(\boldsymbol{r}(t))dt \tag{1.7}$$

其中 $w(t)=\exp\left(-\int_0^t\sigma(\boldsymbol{r}(s))ds\right)\cdot\sigma(\boldsymbol{r}(t))$，$\boldsymbol{r}(t)$ 表示从相机中心打出的光线。要生成最后的彩色图像，还需要一个神经渲染器来处理(1.7)式中的特征图 $\boldsymbol{I_F}$。将这个神经渲染器记为 $\pi_\zeta$，$\zeta$ 代表该模块所有可学习的网络参数，它的具体结构如图 2 所示。

**图 2 神经渲染器结构**

Fig.2 the Structure of Neural Renderer

与 HeadNeRF[14]类似，这个神经渲染器主要由 3 个基本单元组成，每个基本单元均由上采样操作、卷积核尺寸为 3×3 的二维卷积以及 Leaky ReLU 激活函数层构成，通过递归进行上采样操作，实现高效高分辨率的图像合成。特征图每经过一个基本单元都会得到一个分辨率更高的特征图，只要适当的组合这些特征图就可以生成所需要的彩色图像。这里我们借鉴了 Niemeyer 等人在 GIRAFFE[11]中使用的组合方式，用 3×3 的二维卷积核将每一个基本单元作用后的特征图映射到当前分辨率下的 RGB 图像，并通过双线性上采样算子将前一个卷积操作输出的 RGB 图像也采样到当前分辨率下，然后将两个分辨率相同的 RGB 图像逐像素相加，迭代生成目标分辨率下的 RGB 图像。

## 1.3 损失函数

在训练阶段，当给定输入的 RGB 图像时，经过数据处理可以获得其对应的表情隐编码、身份隐编码，以及对应的的相机参数。通过 1.2 节的分析可知，相比传统的 NeRF，本文所提出模型的额外输入只有表情隐编码和身份隐编码，无需引入其他信息如三维几

何等作为监督信号,因此我们的模型仍然继承了 NeRF 的性质，是一个端到端的自监督神经网络。需要注意的是，除了体渲染模块的网络参数之外，神经渲染模块的网络参数也需要在训练阶段得到更新，所有可学习的网络参数在训练阶段是共享的。为了更好地训练模型，损失函数由以下两项构成：

**光度损失** 与(1.3)式相同，在训练阶段，对每一张输入图像，都需要计算渲染误差以优化网络参数。具体来说，通过模型生成的渲染图像的人头部分应该与相应的真实图像的人头部分尽可能相同，可以用公式表达如下：

$$\mathcal{L}_1 = \left\| \boldsymbol{M_h} \oplus \boldsymbol{I_{render}}(\boldsymbol{\beta}, \boldsymbol{z_{id}}, \boldsymbol{P}) - \boldsymbol{I_{GT}} \right\|_1 \quad (1.8)$$

其中 $\boldsymbol{I_{render}}(\boldsymbol{\beta}, \boldsymbol{z_{id}}, \boldsymbol{P})$ 表示在表情隐编码为 $\boldsymbol{\beta}$ 、身份隐编码为 $\boldsymbol{z_{id}}$ 、相机参数为 $\boldsymbol{P}$ 时的模型渲染图像，$\boldsymbol{M_h}$ 为图像上人头区域的掩码，结合哈德马积符号 $\oplus$ 可以将感兴趣的区域限定在人头部分并只在此区域内计算光度损失。

**感知损失** 如果将图像生成任务视为图像转换问题，也就是将输入图像转换为输出图像，模型在训练过程中提取了一些高级特征，那么为了生成高质量的图像，需要定义感知损失[27]如下：

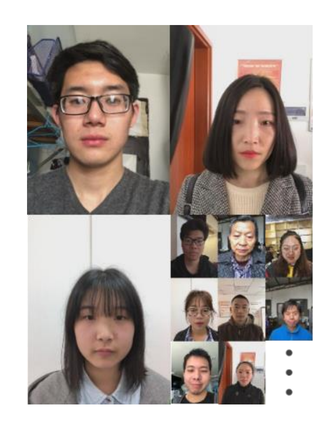
$$\mathcal{L}_2 = \sum_i \left\| \Phi_i(\boldsymbol{I_{render}}(\boldsymbol{\beta}, \boldsymbol{z_{id}}, \boldsymbol{P})) - \Phi_i(\boldsymbol{I_{GT}}) \right\| \quad (1.9)$$

其中 $\Phi_i(*)$ 表示 VGG16[28]网络中第 $i$ 层的激活函数。

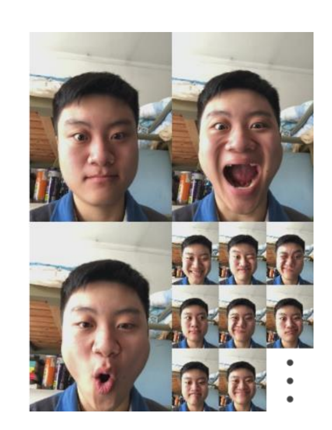
最终的损失函数由光度损失和感知损失加权得到，如下式所示：

$$\mathcal{L} = \mathcal{L}_1 + \lambda \mathcal{L}_2 \quad (1.10)$$

$\lambda$ 为感知损失项 $\mathcal{L}_2$ 的权重系数。



(a)多身份展示



(b) 多表情展示

**图 3 数据集**

Fig.3 Dataset

## 1.4 数据集和数据处理

为了更好地训练模型，我们收集并整理了一个单目动态视频数据集。具体来说，我们采用 iPhone X 手机拍摄了 570 个不同身份的 RGB 视频数据,拍摄对象人种为中国人，在性别、着装、发型上均有不同，每段视频中被拍摄对象的头部转动角度、面部表情都较为丰富。需要特别说明的是，我们的数据中有很多戴眼镜的身份，后续模型对眼镜的拟合能力正是依赖于此。部分数据展示在图 3 中。其中 540 个身份将会用于训练，余下 30 个身份则不会在训练时出现，以此作为测试集评估模型在新身份上的泛化能力。

为了适应模型训练，首先我们使用现有的基于网格的跟踪方法[29]来追踪每个视频中的面部位置，并通过拟合 3DMM 模型[18]得到每一帧的表情系数和头部姿态参数。与 HeadNeRF[14]相同，我们将头部姿态参数作为相应帧的相机外参，这种操作隐含地将每帧的底层几何结构对应到相同的空间位置，减小了相机参数误差对渲染结果造成的影响。其次需要获得身份隐编码，目前开源的准确度最高的人脸识别算法 AdaFace[26]提供了在海量不同身份数据集下训练过的预训练模型，我们将该预训练模型从视频每帧图像中提取的人脸特征信息作为身份隐编码。最后，通过现有的分割算法[30]生成每一帧的头部掩码，以保证损失函数只在头部区域计算。图 4 展示了在单个身份上的数据处理结果，其中头部分割结果由头部掩码作用于 RGB 图像上得到。特别地，从参数化模型的拟合结果来看，数据处理阶段得到的表情系数较为准确的表达了原始 RGB 图像的表情信息，这对模型是否能精确驱动来说非常重要。

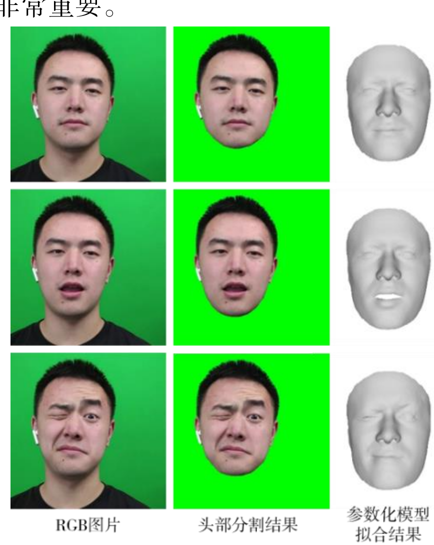



**图 4 数据处理结果**

Fig.4 Dataset Processing Results

经过上述数据处理步骤后，我们得到了由 540 个不同身份的 129552 张人头图像组成的训练集，这些数据被全部打乱按照随机顺序用于模型训练。多身份、多表情的数据集为模型的拟合能力和泛化能力提供了夯实的基础。

## 2 实验结果

### 2.1 实现细节

我们采用 PyTorch 深度学习框架[31]实现本文建立的泛化人头模型，使用 Adam 优化器[32]更新可学习的网络参数。身份隐编码和表情隐编码的维度分别为 $z_{id}\in\mathbb{R}^{512}$，$\beta\in\mathbb{R}^{46}$，(1.10)式中感知损失项的权重系数 $\lambda=10$。文中展示的实验结果均为 batch size 设置为 4 时在 2 个 Testla-V100 GPU 上训练 20 轮的结果，一轮训练有 129552 张图片，20 轮训练总共耗时 70 个小时。

### 2.2 模型评估

#### 2.2.1 解耦控制

在本部分，我们测试了模型对渲染结果的各种语义属性的独立控制能力。如图 5 所示，对于给定的表情隐编码和身份隐编码 $(\beta, z_{id})$，我们可以直接调整相机参数，以连续更改渲染视图的相机位置。特别地，眼镜在不同相机位置下仍然保持了完整且合理的形状。这些新视角合成的渲染结果表明我们的模型具有良好的多视角一致性，尽管没有采用传统 NeRF 的体渲染方式，结合二维神经渲染的渲染模式仍然有效地保留了原始 NeRF 通过位置隐式编码的几何结构。

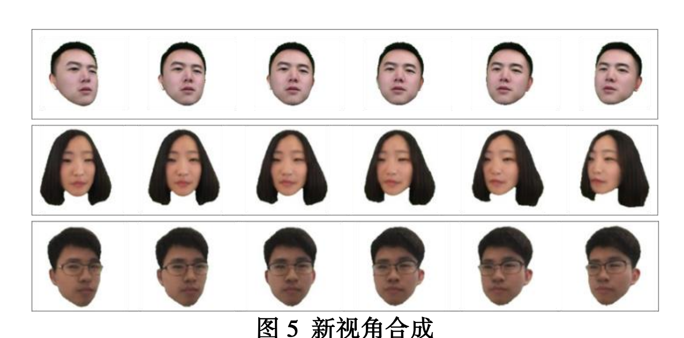


**图 5 新视角合成**

Fig.5 Novel View Synthesis

更进一步地，我们可以利用训练好的模型实现语义控制，独立编辑身份和表情属性。即当身份或者表情两者中任意一个属性给定时，都可以在给定属性不变的情况下实现另一属性的平滑变化。具体地说，当需要编辑身份时，我们在训练集中随机采样两个不同身份的样本，将其中一个视为原身份，另一个视为目标身份，然后在原身份和目标身份的隐编码间进行线性插值得到若干个新的身份隐编码，并分别与原身份的表情编码一同重新渲染人头图像，即可得到身份编辑的渲染结果。同样地，当需要编辑表情时，在训练集中随机采样相同身份的两个不同表情的样本，再对原表情和目标表情进行线性插值并经过重新渲染得到同一身份新表情下的合成图像。如图 6 所示，这种控制变量的插值结果表明，我们的模型能在编辑特定属性的同时维持其他属性不变，有效的解耦了身份和表情的语义信息。




**图 6 语义解耦结果**

Fig.6 Semantic Disentanglement Results

#### 2.2.2 消融实验

**感知损失消融实验** 这个部分我们测试感知损失项对模型渲染结果的影响。对于无感知损失项的模型，它与第 1 节建立的完整模型采用同样的训练策略与训练时长，唯一的区别在于将感知损失项在损失函数中的权重系数设置为 $\lambda=0$。正如图 7 所示，感知损失项显著提高了渲染图像的质量，对细节的展现尤为重要。这里我们特别指出，在图 7 中，保留感知损失除了提高眼睛部分的生成质量外，人脸上的痣和眉毛的毛流感也被很好的渲染出来。

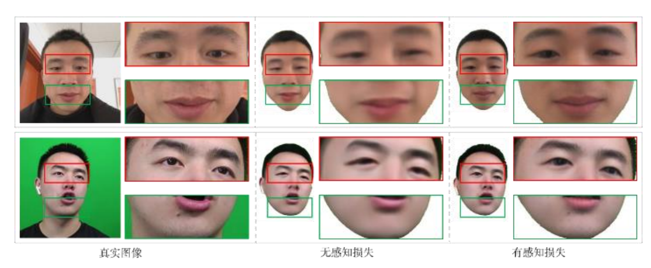



**图 7 感知损失消融实验**

Fig.7 Ablation Study on the Perceptual Loss

**身份编码消融实验** 为了测试由不同方式编码的身份信息对实验结果的影响，我们对此进行了消融实验。这里我们采用两种方式获取身份编码，一种是模型中

所使用的，将人脸识别网络提取的身份特征作为身份隐编码；另一种则是通过(1.4)式拟合人脸参数化模型得到身份系数，将其作为身份隐编码，维度为 $\alpha \in \mathbb{R}^{100}$。两种不同的身份隐编码除了维度不同导致网络输入的维度不同外，其他实验设置全部相同。从图 8 可以看出，参数化模型的身份系数与人脸识别网络提取的身份特征相比，在视觉上明显与真实图像的身份差距更大，并且眼睛部分的渲染细节也更少。为了更加严谨地说明这一问题，我们用多个评价指标定量评估了图 8 中展示的 4 个身份在不同身份编码下的渲染结果，数据如表 1 所示。这里的评价指标 $L_1$，PSNR，SSIM 分别表示真实图像与渲染图像在头部掩码区域内的平均 $L_1$ 范数距离、峰值信噪比( Peak Signal to Noise Ratio, PSNR )以及结构相似性( Structural Similarity, SSIM )，表 1 中更优的结果已加粗显示。

**表 1 不同身份隐编码渲染结果的定量比较**

Table 1 Quantitative Comparison on Different Identity Codes

| 身份隐编码方式 | $L_1$↓ | PSNR↑ | SSIM↑ |
| --- | --- | --- | --- |
| 3DMM 身份系数编码 | 0.077 | 16.9 | 0.952 |
| 人脸识别网络特征编码 | **0.073** | **17.2** | **0.955** |

在这一消融实验中，视觉与数值上的比较结果都强有力地证明了在建立模型时采用人脸识别网络提取的特征作为身份隐编码的正确性与必要性。




**图 8 身份编码消融实验**

Fig.8 Ablation Study on the Identity Code

## 2.3 对比实验

这一小节将评估本文所建立模型的泛化能力，即对于训练集从未出现的新的身份，模型是否仍具有可观的拟合能力。为了有所对比，我们采用同类型的基于神经辐射场的参数化人头方法 HeadNeRF[14]作为比较。

前面 1.4 小节提到，所拍摄的数据中有 30 个身份未参加模型训练，可以用于此处作为测试集。需要注意的是，HeadNeRF 的训练集为 FaceSEIP、FaceScape[33] 和 FFHQ[3]，这三个数据集均由外国人构成，而我们在训练模型时用到的数据都为中国人。由于肤色导致的面部纹理差异，不同人种的训练集会造成模型的渲染效果有所不同。为了公平起见，我们在网上搜集了 30 个不同身份的国外新闻播报视频作为补充测试集。这样测试集由 60 个不同身份构成，其中 30 个为中国人，30 个为外国人。将这 60 个视频按照 1.4 节介绍的数据处理方法构建我们方法的测试集，按照 HeadNeRF 提供的数据处理代码构建对方实验需要的测试集。测试 HeadNeRF 拟合能力的代码是由作者提供的开源代码。




**图 9 泛化能力定性比较结果**

Fig.9 Qualitative Comparison Results for Generalization

图 9 展示了我们的方法与 HeadNeRF 的定性比较结果。从图中可以看出，对于女性的长发，我们的方法有更好的拟合结果。此外，当人脸偏转角度较大时，我们的方法在恢复偏转角度的同时仍然很好的保留了身份信息。更重要的是，正如 HeadNeRF 在其文章中提到的，他们的训练集不曾涵盖头部配件如发卡、眼镜等，因此对于配戴眼镜的拟合对象，HeadNeRF 无

**表 2 泛化能力定量比较结果**

Table 2 Quantitative Comparison Results for Generalization

| | 中国人测试集 | | | 外国人测试集 | | | 全部测试集 | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | $L_1$↓ | PSNR↑ | SSIM↑ | $L_1$↓ | PSNR↑ | SSIM↑ | $L_1$↓ | PSNR↑ | SSIM↑ |
| HeadNeRF | 0.149 | 14.0 | 0.919 | 0.071 | 19.6 | **0.930** | 0.110 | 16.8 | 0.925 |
| Ours | **0.049** | **23.0** | **0.950** | **0.063** | **20.8** | 0.927 | **0.056** | **21.9** | **0.939** |

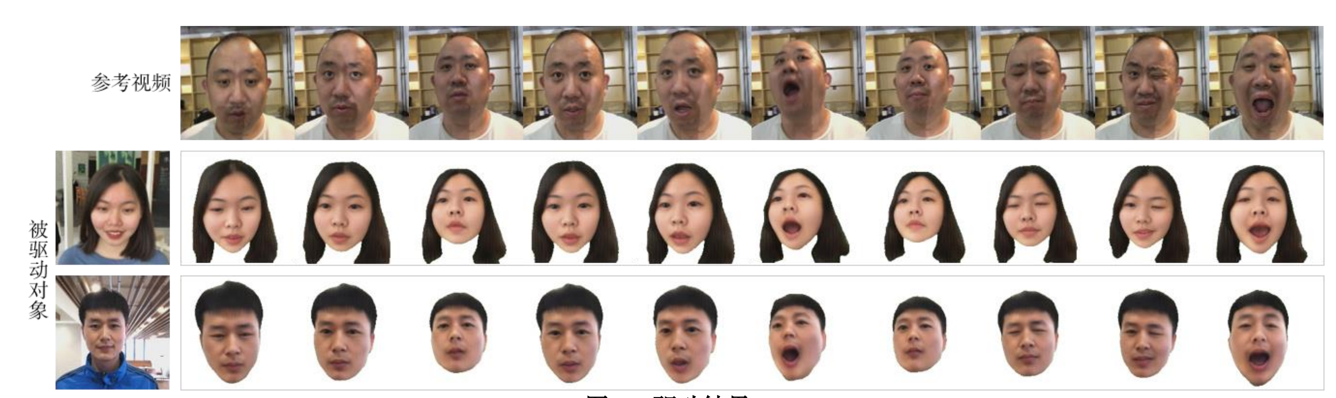


**图 10 驱动结果**
Fig.10 Driven Results

法渲染出眼镜。而我们的方法得益于训练集中有很多配戴眼镜的身份，渲染结果很好的恢复了眼镜形状。

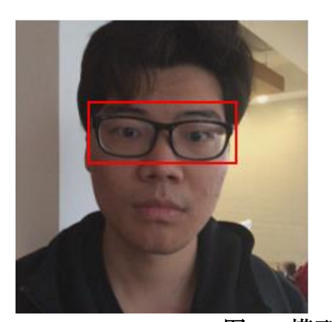
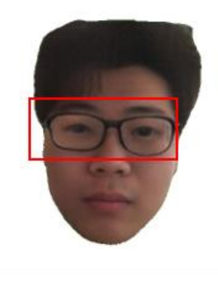
表 2 给出了用 $L_1$ 范数、PSNR、SSIM 三种不同指标评估两种方法在测试集上拟合能力的数值，这里同样计算的是渲染结果与真实图像在头部掩码区域的误差。这些数值结果印证了前文提到的训练集人种不同的问题，我们的方法在中国人的测试集上各项指标明显优于外国人测试集，HeadNeRF 则正好相反。尽管如此，本文提出的方法在外国人测试集上的数值结果也大都优于 HeadNeRF，在结构相似性指标 SSIM 上虽然略低于 HeadNeRF 但相差甚少，可以说达到了相当的水平。定量比较的结果表明我们的模型具有更好的泛化能力。

## 2.4 驱动应用

因为我们建立的模型具有很强的表示能力，可以解耦渲染结果的各种属性，所以它有多种用途，比如新视角合成、表情迁移等。这一节将展示本模型的驱动功能，也就是将参考视频中人物的头部动作与表情在目标人物图像的面部重现。为此，我们需要从参考视频中获取头部姿态和表情隐编码，与目标对象的身份隐编码结合，使用训练好的头部模型生成期望的面部图像序列，再按照对应的时间顺序形成视频，就实现了一个完整的驱动流程。图 10 中展现了部分帧数的驱动结果。值得注意的是，在模型训练阶段参考视频的表情域和姿态域并非与被驱动对象的表情域和姿态域完全相同，因此这里逼真的驱动结果表明模型在表情和姿态上也具有很好的泛化性能。

## 2.5 未来工作

虽然我们的方法建立了一个可解耦的高质量头部泛化模型，但仍然存在一定问题。如图 11 所示，部分渲染结果的眼球部分会出现眼黑占比远大于眼白的现象，导致渲染图像的眼睛不够自然。这是因为我们所采用的训练集是视频数据，而在视频采集过程中拍摄对象的眼珠并未全程顶着相机不动，数据处理过程眼珠的转动与人脸姿态就会耦合在一起，造成模型无法完全学习到眼珠的位置变化。未来可以考虑加入眼球追踪系数或者视线方向等信息进一步缓解这一问题，提高眼球部分的渲染细节，甚至可以做到眼神注视方向的编辑。



**图 11 模型拟合缺点**
Fig.11 Limitation of Fitting

## 3 结束语

本文建立了一个基于 NeRF 的参数化头部模型，它将神经辐射场集成到人脸参数化模型上，并结合人脸识别网络实现了可泛化的头部建模。得益于精心设计的网络结构和损失函数，本模型可以在现代 GPU 上快速渲染高保真头部图像，并且支持更改渲染视角，可以独立编辑生成图像的身份和表情。实验结果表明，我们建立的可泛化人头模型优于目前的相关方法，相信在不久的将来也会为数字人的发展添砖加瓦。

参考文献:

[1] CAO C, WENG Y, ZHOU S, et al. Facewarehouse: A 3d facial expression database for visual computing[J]. IEEE Transactions on Visualization and Computer Graphics, 2014, 20(3): 413-425.
[2] CAO C, CHAI M, WOODFORD O J, et al. Stabilized real-time face tracking via a learned dynamic rigidity prior[J]. ACM Transactions on Graphics, 2018, 37(6): 233.
[3] KARRAS T, LAINE S, AITTALA M, et al. Analyzing and improving the image quality of styleGAN[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 8110-8119.
[4] KARRAS T, LAINE S, AILA T. A style-based generator architecture for generative adversarial networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 4401-4410.
[5] GHOSH P, GUPTA P S, UZIEL R, et al. GIF: Generative interpretable faces[C]//International Conference on 3D Vision (3DV). 2020: 868-878.
[6] DENG Y, YANG J, CHEN D, et al. Disentangled and controllable face image generation via 3d imitative-contrastive learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 5154-5163.
[7] MILDENHALL B, SRINIVASAN P P, TANCIK M, et al. NeRF: Representing scenes as neural radiance fields for view synthesis[C]//Europeon Conference on Computer Vision. 2020.
[8] CHAN E R, LIN C Z, CHAN M A, et al. Efficient geometry-aware 3d generative adversarial networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 16123-16133.
[9] DENG Y, YANG J, XIANG J, et al. GRAM: Generative radiance manifolds for 3d-aware image generation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
[10] GU J, LIU L, WANG P, et al. StyleNeRF: A style-based 3d aware generator for high-resolution image synthesis[C]//International Conference on Learning Representations. 2022.
[11] NIEMEYER M, GEIGER A. GIRAFFE: Representing scenes as compositional generative neural feature fields[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 11453-11464.
[12] SCHWARZ K, LIAO Y, NIEMEYER M, et al. GRAF: Generative radiance fields for 3d-aware image synthesis[C]//In Advances in Neural Information Processing Systems: 33. 2020: 20154-20166.
[13] ZHOU P, XIE L, NI B, et al. CIPS-3D: A 3d-aware generator of GANs based on conditionally-independent pixel Synthesis[M]. arXiv, 2021.
[14] HONG Y, PENG B, XIAO H, et al. HeadNeRF: A real-time NeRF-based parametric head model[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
[15] CHEN A, XU Z, ZHAO F, et al. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 14124-14133.
[16] YU A, YE V, TANCIK M, et al. pixelNeRF: Neural radiance fields from one or few images[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[17] JOHARI M M, LEPOITTEVIN Y, FLEURET F. GeoNeRF: Generalizing NeRF with geometry priors[C]//Proceedings of the IEEE international conference on Computer Vision and Pattern Recognition (CVPR). 2022.
[18] BLANZ V, VETTER T. A morphable model for the synthesis of 3d faces[C]//Siggraph Conference Proceedings, 1999. 1999: 187-194.
[19] AGARWAL S, SNAVELY N, SIMON I, et al. Building rome in a day[C]//2009 IEEE 12th International Conference on Computer Vision (ICCV). 2009: 72-79.
[20] SCHNBERGER J L, ZHENG E, POLLEFEYS M, et al. Pixelwise view selection for unstructured multi-view stereo[J]. Springer, Cham, 2016.
[21] SITZMANN V, ZOLLHFER M, WETZSTEIN G. Scene representation networks: continuous 3d-structure-aware neural scene representations[C]//In Advances in Neural Information Processing Systems. 2019.
[22] CHEN M, ZHANG J, XU X, et al. Geometry-guided progressive NeRF for generalizable and efficient neural human rendering[J]. arXiv e-prints, 2021.
[23] PENG S, ZHANG Y, XU Y, et al. Neural Body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[24] KWON Y, KIM D, CEYLAN D, et al. Neural human performer: learning generalizable radiance fields for human performance rendering[C]. 2021.
[25] PAYSAN P, KNOTHE R, AMBERG B, et al. A 3d face model for pose and illumination invariant face recognition[C]//In IEEE International Conference on Advanced video and signal based surveillance. 296-301.
[26] MINCHUL K, JAIN A K, LIU X. AdaFace: Quality adaptive margin for face recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
[27] JOHNSON J, ALAHI A, FEI-FEI L. Perceptual losses for real-time style transfer and super-resolution[C]//Proceedings of the IEEE/CVF European Conference on Computer Vision. 2016: 694-711.
[28] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. Computer Science, 2014.
[29] GUO Y, ZHANG J, CAI J, et al. CNN-based real-time dense face reconstruction with inverse-rendered photo-realistic face images[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2018, 41(6): 1294-1307.
[30] KE Z, SUN J, LI K, et al. MODNet: Real-time trimap-free portrait matting via objective decomposition[C]//Association for the Advancement of Artificial Intelligence. 2022.
[31] PASZKE A, GROSS S, MASSA F, et al. PyTorch: An imperative style, high-performance deep learning library[C]//In Advances in Neural Information Processing Systems. 2019.
[32] KINGMA D P, BA J. Adam: A method for stochastic optimization[C]//International Conference on Learning Representations. 2014.

[33] YANG H, ZHU H, WANG Y, et al. FaceScape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 601-610.